CITO Research
Advancing the craft of technology leadership

Optimized Data Engineering: A Crucial Challenge for the Internet of Everything Era

COSPONSORED BY

MAPR®

CISCO™

CONTENTS

Introduction

How often do $19 trillion opportunities come around? The question isn't outlandish. Whether businesses realize it or not, we are undergoing one of the most monumental shifts in the way we work and live since the advent of the Internet. Comparisons to the invention of the printing press and Gutenberg's Bible might be slightly exaggerated—but only slightly. The era of the Internet of Everything (IoE) is upon us and IoE will inevitably change the way companies operate.

But too much of the discussion about IoE glosses over the skills and capabilities required of data engineering. Too often we hear the following narrative:

1. IoE will bring a deluge of connected devices that can tell us many things.
2. The devices will be connected to each other, to people, and with processes.
3. A flood of data will arrive from all of these connections.
4. This will provide businesses with the potential to better run their businesses and make their customers happier.

Exploration of activity in 3 and 4 is limited. Massive amounts of data must be landed, cleaned, organized, and analyzed. New types of applications must be created. New processes, products, and services will emerge, based on finding meaning in the flood of data. The data is so different from what we have now that old paradigms just won't do. What IoE brings about is not an ETL problem or a data warehousing problem, but a problem of data engineering.

What IoE brings about is not an ETL problem or a data warehousing problem, but a problem of data engineering

The goal of this paper is to describe the data engineering challenges brought on by IoE and suggest how companies will have to increase their skills and capabilities.

The Many Transformations of the Internet of Everything

IoE affects so much, which is why it gets so much attention. Our analysis is limited to focusing on the skills, capabilities, and computing infrastructure needed to handle the flood of data, but IoE will also dramatically impact the following areas:

Product definition: How will smart and connected devices change the nature of products? What products will become services? What new services are possible? What data do we need to acquire?

Analytics: What new questions can we answer? What new models can we build? What analytics skills must we develop? What can we predict about the future? What kind of analytics can help in real time?

Process design: How can a more accurate model of the world drive business? How can we change relationships between people running our businesses and improve our relationship with our customers?

Application development: How will IoE applications be structured? When should they be centralized or distributed? What are the right tools to build them? How can data flow between them?

The key to answering these questions depends on improving capabilities in data engineering. In the past, data engineering was a relatively stable process that didn't have to scale and handled a narrow range of data in a one-way flow. The extract, transform, and load (ETL) process used for loading a data warehouse is the canonical example.

In the world of IoE, data engineering means handling large volumes of many kinds of data, much of it in real time. That data cannot be easily cleaned or understood without machine learning and analytics. Massive amounts of data may need to be processed in a short time, as it arrives. The way that data is used may not be determined until long after it is stored. The structure and meaning of the data may change. The data may flow to and from many repositories and applications.

While it is possible to do a big data proof of concept using almost any type of technology, the ability to process IoE data at scale requires an integrated software, hardware, and networking approach. Without such a system, costs or delays can quickly become unbearable.

Understanding IoE

While the Internet of Things (IoT) market has garnered significant momentum recently, some companies may not be as familiar with IoE. IoE and IoT are not the same: IoT is one of three types of connections that make up IoE. IoT refers to the data generated by increasingly self-aware devices—machine-to-machine technology. This includes light bulbs that alert users to their remaining life, connected household appliances that can be monitored from anywhere, and security alarms and locks that can be controlled remotely. By 2020, Cisco estimates that there will be 10 billion connected things shipped annually and a total of 50 billion things in use. Increasingly, our devices will talk among themselves and monitor their own performance.

But IoE accounts not just for IoT, but also for other ways our lives will be connected. In addition to machine-to-machine connections, IoE incorporates people-to-people and people-to-machine connections. People-to-people connections are obvious, including social media as well as collaboration activities. People-to-machine connections depict the way in which we will interact with connected devices in the future. Our responses to connected devices will influence how IoE shapes our lives as much as the machines themselves.

By 2020, Cisco estimates that there will be 10 billion connected things shipped annually and a total of 50 billion things in use

Yet it's all three connections interacting that present the greatest opportunity. Consider a clothing retailer comparing in-store stock with what customers are saying on social media to ensure that every store has a sufficient quantity of the most popular t-shirts.

The Internet of Things: One of Three Types of IoE Connections

Internet of Things:

Device-to-device, or machine-to-machine

Internet of Everything:

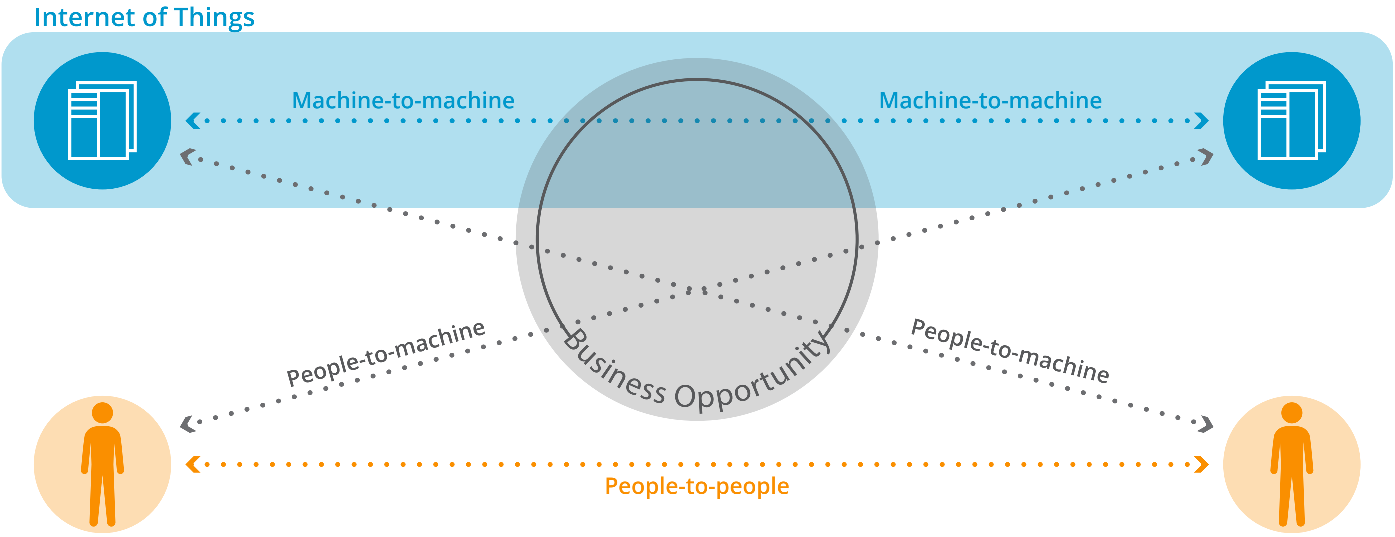
- Machine-to-machine
- People-to-people
- People-to-machine

The IoE market opportunity is leading companies to focus even more on data-driven growth and operations. Gartner estimates that 80–90% of enterprises are considering big data and data-driven strategies. The reason is simple: data-driven enterprises outperform industry peers by up to 6% and are up to 26% more profitable, according to research from the MIT Sloan Center for Digital Business.

Savvy businesses will recognize the opportunity IoE represents. They will be able to gain key intelligence about products and the way customers respond to them. The ability to have everything in a production chain complete with embedded intelligence means better real-time feedback about how that chain is functioning.

Yet many businesses are not prepared from an architecture standpoint to handle the data demands of IoE and risk not being able to take advantage of IoE's potential. Competitive advantage goes to those who use IoE data to better understand customers and business processes and find ways to create value. This means combining skills for data engineering and analytics with the ability to process the data in time for the insights to matter.

Internet of Everything



IoE: A Data Engineering and Business Quandary

IoE will generate more data than we have ever seen. When all things and people are connected at all times, they produce constant data streams, in varying formats. Existing data analytics, warehousing, and processing systems were not designed to handle this type of volume or variation. Companies have to adjust their infrastructure in order to put IoE data to good use.

But there are organizational dynamics at play as well. Line-of-business executives are under pressure to deliver innovations that IoE enables. These executives come to IT and say, “please make this happen.” At first, it won’t be that difficult to move the initial volume of data around, but then, if proof-of-concept projects succeed, data will arrive in massive volumes resulting in the following challenges:

- Ensuring infrastructure is designed for maximum performance at the lowest total cost of ownership as the demand scales
- Ensuring data processing clusters are adequately provisioned with both storage and computing capacity
- Minimizing the time required for creating and tearing down nodes for large jobs or experimentation
- Effectively managing the growing infrastructure to deploy new resources quickly and manage them efficiently
- Optimizing the network to deliver the data efficiently, where and when it is needed
- Running everything in a way that supports enterprise IT requirements, such as compliance, data protection, security, and disaster recovery

The problem is that you can’t design an infrastructure in advance to meet your IoE needs. Even if you could, such an infrastructure would have to be adaptable. And remember, adaptability isn’t enough; scale is also required. It is imperative to select a flexible infrastructure that can cost-effectively support unprecedented scalability.

IoE requires a flexible infrastructure that supports high scalability

Once you have a flexible foundation, lots of design challenges remain. Big data repositories will sit alongside existing data warehouses, creating a type of data supply chain. Applications, mobile devices, and IoE devices will both consume and provide data to this supply chain. It will be tempting to create new data silos for a pressing use case, but the last thing you need is even more silos.

Additionally, in order to handle IoE data, companies must adopt new data and analytic capabilities. Companies have traditionally relied on data warehouses. This worked for structured data, of reasonable volume. But relying on data warehouses to handle big data from IoE, which is often unstructured and in immense quantities, is prohibitively expensive and inefficient. Thus, as the basis of your big data architecture, you need a platform that can handle all types and volumes of big data.

One key aspect of the data engineering infrastructure will be a Hadoop cluster that can be adaptable, highly performant, and integrate easily with the rest of your computing infrastructure.

Hadoop needs to run on a powerful platform and support the right data engineering and applications to keep pace with IoE data produced by all those devices and people. An IoE data-engineering solution must offer upstream and downstream support to process the high-speed write-intensive operations with constant updates that are at the core of IoE data generation and usage.

Tomorrow's Data Analytics

IoE will require a conceptual shift in data analytics. Traditional thinking about analytics was based on a centralized process and a single repository. Data was collected and stored in a central data warehouse, then prepared for use by data analysts who search it within set parameters, extracting takeaways that inform the company's decision making. In this process, data served the analytics. Data was put in one place, where it was munged and transformed. It was very top-down and deliberate and, for many companies, highly successful.

That style of analytics architecture will not work in the IoE era. Putting IoE data to use is predicated on the understanding that analytics no longer involves a few major decisions that guide the course of the entire business over a long-term horizon. Rather, in the IoE era, companies use data constantly, with ongoing small and rapid decisions that enable real-time course adjustments.

Analytics is no longer a top-down process. In fact, with IoE, analysts may not even be the decision-makers at all times. Machines may start to make many of the adjustments and decisions based on the data they're generating, requiring compute at the edge. Companies with a comprehensive big data environment that processes both data at rest and data in motion will be far more efficient and productive.

As a result, ETL will no longer entail collecting data from disparate places, moving it to a centralized hub, and then performing the needed transformations. That system is already archaic. With IoE, a company needs run-time ETL. It is critical to have an architecture that addresses the time value of data and that is capable of offering real-time insight when needed.

One key for companies as they plan their data infrastructure is to create a system that can process data in motion as well as data at rest. Data in motion, such as security camera footage that can lead businesses to make decisions in real-time, needs a different level of support. But key insights—likely the most important ones—will be found using large quantities of data at rest.

The Real World

IoE transformations are already occurring in many industries. A major wind turbine company has every turbine connected and churning out usage data in real time. Because wind is so unpredictable, it's hard to maximize the power turbines can produce.

The company's IoE architecture enables responding to data in real time. Based on wind changes, turbines make constant adjustments to extract the greatest amount of energy. This is the type of real-time IoE data feedback loop that the right infrastructure can create.

IoE transformations are already occurring in many industries

The challenges inherent in creating a platform for data engineering for IoE represent a new frontier for IT and demand a new infrastructure. At the leading Internet and big data companies, years of work have gone into creating purpose-built systems to handle all the data. Hadoop itself is the result of some of this innovation, but so are systems built on top of Hadoop, such as HBase, Hive, Spark, and Storm. Graph databases, application development environments such as Cascading, and dozens of other technologies are used to perform data engineering that supports analytics and application development.

The challenge now is how to create a data-engineering infrastructure the enterprise needs. Few companies outside Silicon Valley will want to knit together a scalable, manageable, and reliable data-engineering infrastructure from raw open source. Remember, big data experts are also cluster experts, creating, configuring, and managing large clusters of computers. This type of computing is new to most enterprise IT departments, which don't want to build everything from scratch.

Cisco and MapR for Big Data Analytics

Cisco and MapR have combined forces to create a highly efficient and high performance infrastructure platform for IoE data engineering and analytics.

The objective of Cisco and MapR's partnership is to productize from top to bottom the hardware, software, and networking capability needed to create a data-engineering infrastructure that has the desired properties described so far.

The offering is founded on Cisco's Unified Computing System (UCS) Common Platform Architecture (CPA) for Big Data, which unites integrated computing, networking, and storage resources. Cisco's Application Centric Infrastructure (ACI) builds on this foundation by providing network and application programmability to optimize performance of big data workloads. Here are a few of the key dimensions of the solution.

Scalability. Cisco UCS provides the ability to scale clusters to up to 10,000 systems, with ever greater efficiencies. UCS systems integrate computing, networking, and storage resources that allow you to scale to process even the largest data clusters. Cisco UCS enables LAN, SAN, and management traffic to travel over the same unified fabric cable, a feature called SingleConnect, which significantly reduces both complexity and cost even as data volumes grow. As applications scale, customers see a 66% reduction in switch ports and cabling over traditional solutions. Using MapR's Distributed NameNode HA architecture as the foundation offers greatly improved scalability over traditional Hadoop configurations with a Primary NameNode.

As applications scale, customers see a 66% reduction in switch ports and cabling over traditional solutions

Manageability. Cisco's Unified Computing System (UCS) enables IT to run clusters of compute and network and automatically deploy, manage, scale, and reconfigure them using software. The MapR Control System (MCS) offers administrators a single screen for configuring, monitoring, and managing their clusters.

Performance. The Cisco UCS CPA for Big Data solution is specifically engineered to handle the most demanding big data workloads. MapR complements the Cisco architecture, offering the most enterprise-grade Hadoop distribution with an enhanced data platform for extreme reliability and ease of management. In addition, MapR's distribution is faster than any other; running on the Cisco UCS CPA for Big Data solution, MapR set the MinuteSort record, sorting 15 billion records in 59 seconds. As you begin to run more big data workloads and different workload types, ACI's programmability also provides powerful performance advantages. For instance, you can use Cisco ACI's network awareness to gain insight into network congestion caused by a mixed workload environment, and then apply policies related to packet prioritization and load balancing that can dramatically improve network performance by as much as 50 to 100 percent, or more.

Together, Cisco and MapR allow companies to create a more easily manageable big data system with robust scalability as data loads increase. With UCS, a company's IT department can handle pools of servers as a single entity, which increases flexibility. Automated management speeds server deployment, reducing provisioning time by 84% when compared to traditional systems. And Cisco provides a unified fabric converging the three different types of network traffic—LAN, SAN, and management traffic—into a single data storage and management interface. Cisco's streamlined solution uses a single pair of fabric Interconnects for up to 160 servers, which replaces potentially dozens of LAN, SAN, and management switches that can be scaled to thousands of servers and managed using UCS Central. Not only is this highly efficient, but it saves money through using far fewer port licenses.

Using the MapR and Cisco infrastructure, companies can dramatically reduce their TCO even as they tackle ever more big data at unprecedented network speeds

From its inception, MapR has been a leader in architectural innovations to make Hadoop easier to use and more dependable. These innovations range from self-healing of critical services, point-in-time data-recovery snapshots and frequent enhancements that help prevent outages. MapR also overcomes the limit on the number of files that can be stored in HDFS and can store up to a trillion files. This is a critical requirement for IoE, where lots of small files need to be managed and ingested.

Using the MapR and Cisco infrastructure, companies can dramatically reduce their TCO even as they tackle ever more big data at unprecedented network speeds.

Conclusion

A $19 trillion opportunity doesn't come along every day. Lines of business will look to IT to help them capitalize on this opportunity. IT will need an enterprise-grade architecture that leverages its existing data engineering and takes it to the next level to handle massive and ever-increasing amounts of real-time big data from IoE. That architecture will need scalability and the ability to be reconfigured to support new use cases. The partnership between Cisco and MapR provides just such an infrastructure.

The data engineering challenges inherent in IoE are profound. Companies need to ensure they have adopted the right IoE architecture to take advantage of the real-time insights IoE data provides. Once they've optimized their big data engineering, every aspect of IoE data process, from storage to analytics, becomes easier and more useful.

Find out more about IoT and IoE

To learn how MapR and Cisco work together to address the crucial challenges of big data, IoT, and IoE, visit www.mapr.com/cisco

Cisco (NASDAQ: CSCO) is the worldwide leader in IT that helps companies seize the opportunities of tomorrow by proving that amazing things can happen when you connect the previously unconnected.

MapR delivers on the promise of Hadoop with a proven, enterprise-grade platform that supports a broad set of mission-critical and real-time production uses. MapR brings unprecedented dependability, ease-of-use and world-record speed to Hadoop, NoSQL, database and streaming applications in one unified distribution for Hadoop.

The paper was authored by CITO Research and cosponsored by MapR and Cisco.

CITO Research

CITO Research is a source of news, analysis, research and knowledge for CIOs, CTOs and other IT and business professionals. CITO Research engages in a dialogue with its audience to capture technology trends that are harvested, analyzed and communicated in a sophisticated way to help practitioners solve difficult business problems.

Visit us at http://www.citoresearch.com